I HEREBY CERTIFY THAT THIS CORRESPONDENCE IS BEING DEPOSITED WITH THE UNITED STATES POSTAL SERVICE AS EXPRESS MAIL IN AN ENVELOPE ADDRESSED TO: ASSISTANT COMMISSIONER FOR PATENTS, WASHINGTON, D.C. 20231. THE APPLICANT AND/OR ATTORNEY REQUESTS THE DATE OF DEPOSIT AS THE FILING DATE.

Express Mail No: ER408660433US

Date of Deposit:: January 15, 2004

Name of Person Making Deposit: Karen Cinq-Mars

Signature: Karen Cinq-Mars 1/15/04

APPLICATION

FOR

UNITED STATES LETTERS PATENT

APPLICANT:
John E. Barth, et al.

FOR:
CONCURRENT REFRESH MODE WITH DISTRIBUTED ROW ADDRESS COUNTERS IN AN EMBEDDED DRAM

DOCKET:
FIS920030409US1

INTERNATIONAL BUSINESS MACHINES CORPORATION
ARMONK, NEW YORK 10504

Concurrent Refresh Mode with Distributed Row Address Counters in an Embedded DRAM

Background of the invention

This invention is generally related to an embedded dynamic random access memory (embedded DRAM), and more particularly, to a concurrent refresh mode and design which employs distributed row address counters integrated in each DRAM.

Improvements in semiconductor technology have enabled the design of processors having performance exceeding 1 Giga Hz. However, the system performance is often constrained by the performance of its memory. The presence of this drawback has created a potentially strong demand for high performance embedded DRAMs to help the processor achieve the necessary speed. For 90nm technology generations and beyond, it is difficult to reduce the cell size and still improve the array access transistor performance of the embedded DRAM. This is true because the transistor threshold voltage cannot be reduced when device leakage is present. Yet, the operation voltage must be reduced to guarantee the device reliability and logic process compatibility. These considerations have created a fundamental shift from a data retention driven design to a memory availability driven design by utilizing a high performance logic device as a memory cell.

Referring to Fig. 1, there are shown simulated sensing signals illustrating the aforementioned assertions. In the plot referenced (A), a conventional array device 52A (not shown) having a 2.5V wordline boosted voltage (VPP) and supported by a long bitline coupling 256 cells (256b/BL) is compared to the plot referenced (B), wherein a logic array device 22A (not shown) powered by a 1.5V VPP is supported by a short bitline coupled to 64cells. The sensing signal is extracted by changing the signal development time (tSIG), i.e., the time to develop a signal on the bitline when a wordline is activated over a random access cycle time (tRC) in a grounded sensing scheme. As the

signal development time increases, an excess of charge in the cell is transferred to the bitline, increasing the sensing signal. However, when tSIG is incremented by more than about 40% of tRC, the voltage is not adequately written back to the cell, resulting in a smaller signal. Since the logic array device 22A is enabled by approximately 30% more current than that applicable to the corresponding array device 52A, approximately 80mV of the sensing voltage can be achieved even for a 3.2ns random access cycle time. However, employing a logic array device 22A requires a reduction in the data retention time to a value as small as 64μs. The shorter retention time greatly reduces the memory availability, particularly for a large density memory because all the memory cells need to be refreshed within a given retention time to maintain the data bits. By way of example, a 4Mb memory having 8K wordlines requires 8K refresh cycles within 64μs. This, in turn, requires at least one refresh command every 8ns, resulting in the memory being unavailable for an 8ns random cycle memory. In order to overcome the memory availability problem in short retention DRAMs, a concurrent refresh mode is typically used, as described, e.g., in U.S. Patent No. 4,185,323 issued to Johnson et al.

Fig. 2 is a block representation of a semiconductor memory chip 200 consisting of a plurality of DRAM memory banks 210. Each memory bank 210 consists of a plurality of DRAM memory cells (not shown) arranged in a two-dimensional matrix configuration, well known in the art, and which, accordingly, will not be discussed further. Once a memory access operation (read, write, or refresh operation) is initiated in a DRAM bank 210 (e.g., 210i), the DRAM bank (e.g., 210i) becomes unavailable for a random access cycle time tRC. During the memory access operation of the DRAM bank (e.g., 210i), other DRAM banks (e.g., 210j) can be simultaneously refreshed. Thus, the memory availability greatly improves by concurrently performing a refresh operation while enabling a memory access operation. There are two known methods for enabling a concurrent refresh mode in a semiconductor memory, the details of which will be discussed hereinafter, as explained hereinafter with reference to Figs. 3 and 4.

Fig. 3 illustrates a first method to enable a concurrent refresh mode in a conventional static random access memory (SRAM) buffer. Details of this approach are

described, e.g., in U.S. Patent No. 5,999,474, issued to Leung et al. Semiconductor memory chip 300 consists of a plurality of DRAM banks (310DRAM), each consisting of a plurality of memory cells arranged in a two dimensional array configuration. Accessing bank 310DRAM-310j while concurrently refreshing at least one other bank 310DRAM - 310k is possible as long as the accessed bank and the refreshed bank differ from one another. This allows a plurality of cells (330k) supported by the corresponding wordline 320k in DRAM bank (310k) to be refreshed while accessing a plurality of cells (330j) supported by the corresponding wordline (320j) in the DRAM bank (310j). However, if array 310j is continuously addressed, some memory cells within the same array 310j will not be refreshed altogether since array 310j is continuously busy due to an uninterrupted memory access operation. This precludes performing a refresh operation of some memory cells in the same array (310j).

In order to overcome this problem, memory chip 300 is enhanced by adding an SRAM bank (310SRAM), featuring a dual port function that allows receiving and transferring data within a clock cycle. The access operation of the DRAM banks (310DRAM) and SRAM (310SRAM) are controlled by the TAG memory (310TAG), while the memory access of the memory chip 300 is enabled by a read or write command (not shown), a bank address (XBADD), and a word address (XWADD), wherein XBADD and XWADD identify one of the DRAM banks (310DRAM) and the appropriate wordline within the selected DRAM banks. When the memory access is enabled, wordline (320TAG) in TAG memory (310TAG) and wordline (320s) in SRAM bank (310SRAM) are activated by decoding the word address (XWADD). This enables reading out data in the memory cells (330TAG) in within the TAG memory (310TAG) and data in the memory cell (330s) within the SRAM buffer (310SRAM). The read data bits (330TAG) of TAG memory (310TAG) defines the bank address (TBADD) which, in turn, identifies the corresponding DRAM bank for the data bits (330s) currently read from the SRAM buffer (310SRAM). When TBADD coincides with the bank address input (XBADD), the data bits (330s) are the ones that are requested by the memory access command, since the data bits (330s) were previously copied from the corresponding DRAM bank to the SRAM buffer (310SRAM). Therefore, no DRAM

bank access is necessary, and the read data bits from the SARM buffer (310SRAM) are read out from the XDATA pins. On the other hand, if TBADD differs from the bank address inputs (XBADD), the TAG memory (310TAG) controls the DRAM banks (310DRAM) as follows.

Assuming that TADD identifies DRAM bank (310i), then, the data bits (330s) in the SRAM buffer (310SRAM) are stored back in DRAM bank (310i), where the wordline 320i is same as the wordline address of 320s (Direct Mapping). This allows data bits to be transferred from the SRAM memory cells (330s) to the DRAM memory cells (330i). Concurrent with the bank address input (XBADD), the corresponding DRAM bank (310j) is activated for a read operation. Then, the cells' data bits (330j) in the corresponding DRAM bank (310j) are read out, where wordline 320j coincides with the wordline address of 320s (Direct Mapping). They are read out from the XDATA pins. The cells' data bits (330j) are also stored in the cells (330s) of the SRAM buffer (310SRAM). TBADD is therefore updated to identify the DRAM bank 310j for a future memory access command. For a subsequent same addressing pattern (i.e. 330j), data bits are read out or written into the SRAM buffer (310SRAM), enabling a refresh operation of the memory cells even when only one array (i.e. 330j) is continuously addressed. This is possible since, eventually, the data bits in the array will be copied to the SRAM array, refreshing the array without infringing on any violations.

This concurrent refresh approach, however, has several drawbacks. First, it requires an SRAM array (310SARM), which is significantly larger. Secondly, on account of the TAG management, the logic becomes more complex which, in turn, slows down the latency of the memory access. Finally, this methodology is not appropriate for multi-bank memories since the memory bank becomes unavailable during a refresh operation within a given DRAM bank cycle (tRC). Multi-bank memory chips require addressing any bank that need to be addressed during each bank-to-bank access cycle (tRRD) -- which is shorter than tRC -- making it impossible to enable a refresh operation when a tRC cycle is required.

Fig. 4 illustrates the second method that enables a simultaneous refresh by utilizing a concurrent function for the DRAM. Semiconductor memory chip 400 consists of a plurality of DRAM banks 410 (410i through 410j), each of which is controlled by the corresponding address and command ports (420i through 420j). Therefore, any two or more banks can be activated concurrently. By way of the concurrent function, memory bank 410i remains in the read mode while still enabling a refresh operation for memory bank 410j. However, this approach requires a complex refresh system management to avoid a bank access contention caused by the concurrent function. While avoiding a bank access contention by the concurrent function, handling the refresh addresses in each array at the system level is highly complex since the address TAG for the refreshed memories for all the banks needs to be independently managed. As a result, employing the concurrent function for a simultaneous refresh requires significant system modifications.

Objects and Summary of the Invention

Accordingly, it is an object of the invention to provide a concurrent refresh operation to an embedded DRAM to improve the memory availability.

It is another object of the invention to provide a concurrent refresh operation to an embedded DRAM without resorting to using a SARM buffer.

It is still another object of the invention to provide a concurrent refresh operation to an embedded DRAM in order to simplify the design of a memory system.

It is further object of the invention to enable a concurrent refresh operation to an DRAM embedded in a multi-bank memory system.

It is still a further object of the invention to enable a concurrent refresh operation to an embedded DRAM resorting to only a refresh bank selection.

This invention describes a concurrent refresh mode, wherein the embedded DRAM enables a simultaneous memory access and refreshes the operations by way of a

simple system modification. The concurrent refresh mode is realized by allowing the unselected memory array to be refreshed only by a refresh bank select port. Unlike conventional approaches, macro row address counters integrated in each bank track the wordline address within the corresponding banks. This greatly reduces the complexity of managing the refresh address in a concurrent refresh mode, since the in-macro refresh counter in each bank maintains independently the wordline refreshed. The system improvement that employs this concurrent refresh method is achieved solely by managing the bank access contention. The present invention is particularly advantageous for a multi-bank system having a short retention DRAM, since the refresh management can be integrated within the existing multi-bank management system. As long as the bank contention is managed, 100% of memory availability can be realized.

In another aspect of the present invention, there is provided a semiconductor memory consisting of two or more memory arrays, wherein each of the two arrays is coupled to a row address counter to generate a first word address within each array when a refresh command is given, while enabling at least one more array to be in a memory access mode.

In yet another aspect of the invention, there is provided a semiconductor memory that includes: i) a plurality of memory arrays, each of which comprises a plurality of memory cells arranged in a matrix and controlled by a row address counter uniquely assigned to each of the memory arrays, the row address counter generating a first word address; and ii) means for enabling a refresh operation in the memory cells, the memory cells being identified by the first word address when a refresh command is issued to a corresponding memory array.

Brief Description of the Drawings

The accompanying drawings, which are incorporated in and which constitute part of the specification, illustrate presently preferred embodiments of the invention and, together with the general description given above and the detailed description of the preferred embodiments given below, serve to explain the principles of the invention.

Fig. 1 shows two plots representing simulated sensing signals generated from two 256b/BL and 64b/BL DRAM arrays that illustrate the conventional shift from a data retention driven design to a memory availability driven design.

Fig. 2 shows a block diagram representing a multi-bank DRAM memory device to illustrate how the memory availability improves by a applying a prior art concurrent performance of a refresh operation while enabling the memory access

Fig. 3 illustrates a prior art method to enable a concurrent refresh mode in memory chip 300 which has been is enhanced by adding an SRAM, and which features a dual port function to receive and transfer data within one clock cycle.

Fig. 4 is shows another conventional method that enables a concurrent refresh mode by utilizing a simultaneous function for a DRAM, wherein by way of the simultaneous function, the memory bank remains in the read mode while still enabling a refresh operation for the memory bank.

Fig. 5 illustrates a memory architecture applicable to a concurrent refresh mode with distributed row address counters, in accordance with the present invention.

Fig. 6 shows a transistor level schematic of the row address counter integrated in each bank of the DRAM, according to present invention.

Fig. 7 shows a detailed bank architecture, consisting of a core, row address and switching elements, according to present invention.

Detailed Description of the Invention

Referring now to Fig 5, there is shown a memory architecture provided with the inventive concurrent refresh mode with distributed row address counters. The present

embodiment assumes an embedded DRAM macro. However the invention is also applicable to a stand-alone DRAM.

The DRAM macro employs a flexible multi-bank protocol having 16 independent bank select ports $BSEL_{0\text{-}15}$, each controlling a corresponding array of $BANK_{0\text{-}15}$. Optionally, $BSEL_{0\text{-}15}$ may be implemented as a four bit bank address vector that identifies one array of $BANK_{0\text{-}15}$. Unlike conventional DRAM macros, it further includes 16 refresh bank select ports $RBSEL_{0\text{-}15}$, each controlling the corresponding array as a bank independently from the memory access operation. The basic concept of the present invention is not a concurrent refresh mode, but the introduction of distributed row address counters integrated in each bank to achieve a greater simplification of the refresh management at the system level.

Each array includes a row address counter RAC (e.g., 520) that identifies the word address $WRAC_{0\text{-}6}$ for a concurrent refresh mode. Each array further includes a switch 530 to selectively couple either the word address $WADD_{0\text{-}6}$ or the word address $WRAC_{0\text{-}6}$ to the row decoder (not shown) of the array bank (BANK). For a memory access operation, the word address $WADD_{0\text{-}6}$ issued by bank select signal BSEL is coupled to the row decoder (not shown) in array 510 via switch 530. This allows the corresponding wordline (not shown) in the array 510 to be activated according to the word address $WADD_{0\text{-}6}$. On the other hand, when a bank refresh command RBSEL is issued, the word address ($WRAC_{0\text{-}6}$) from counter RAC 520 is coupled to the row decoder (not shown) in the array by way of switch 530. This allows the corresponding wordline (not shown) in the array 510 to be activated according to the word address ($WRAC_{0\text{-}6}$) in order to refresh the corresponding memory cells. By managing BSEL (i.e., $BSEL_0$), and RBESL (i.e., $RBSEL_0$), a memory access of an array (i.e., $BANK_0$) is enabled, while concurrently enabling a refresh operation for another array (i.e., $BANK_{14}$). Because the RAC counter is integrated in each bank, the wordline activation within the selected refresh bank (i.e. $BANK_0$) is internally managed. This greatly simplifies the system designs.

As previously mentioned, avoiding a bank contention is a well known design practice for a multi-bank memory system. Assuming that each array consists of 128 wordlines, data will be maintained as long as 128 refresh commands are issued for each bank within the retention time. This results in an almost total memory utilization by properly managing RBSEL. In a multi-bank system, the bank may be activated in a staggered manner for each bank-to-bank access time cycle (tRRD), while concurrently refreshing other arrays for each tRRD. Access contention between the accessed bank and the refreshed bank can be avoided as long as the activation of the same bank is longer than the random access cycle time (tRC). More particularly, activating a subsequent bank (BANKn) either by BSELn and RBSELn must be longer than tRC. This management is much simpler than the existing concurrent refresh management with corresponding address and command ports. No refresh address management for the selected refresh bank is necessary. Optionally, two or more memory arrays may be simultaneously refreshed while enabling the memory access for each clock cycle. This is advantageously realized by simultaneously activating a plurality of the refresh bank control signals $RBSEL_{0\text{-}15}$. The distributed RAC counter approach requires controlling RBSEL for the corresponding refreshed bank without establishing communication between the address port and the respective control circuitry. This results in a current saving as large as 10mA assuming a seven address bus transition, each having a 1.5pF capacitance operating at 1GHz frequency. The current saving advantage is further improved as the memory speed and density increase.

Figure 6 is a transistor level schematic of the row address counter 520 integrated in each bank (BANK). The row address counter includes seven counter logic elements, 610_0 to 610_6, each generating a corresponding address bits $WRAC_{0\text{-}6}$. Each counter logic element (e.g., 610_0) consists of two CMOS pass gates 622 and 624, two CMOS latches 626 and 628, and two inverters 620 and 630. The refresh enable signal RBSEL shown in Fig. 5 is coupled to the NMOS gate for CMOS pass gate 622, and PMOS gate for CMOS pass gate 624. RBSEL is inverted by inverter 620 and coupled to the PMOS gate of CMOS pass gate 622 and NMOS gate of CMOS pass gate 624. Therefore, as long as signal RBSEL is at low, CMOS pass gate 624 couples node N2 to node N3, and

subsequently to node N4. CMOS pass gate 622 remains off, isolating node N0 from N4. The output of $WRAC_0$ (coupling to N4) from the counter logic element 610_0 therefore follows node N1. When the signal RBSEL switches to high to enable a concurrent refresh mode, CMOS pass gates 622 and 624 switch on and off, respectively. $WRAC_0$ maintains the state at the original value set by CMOS latch 628. The state of node N1 is flipped by coupling node N0 to node N1, (Note: because node N0 is in the inverted state from $WRAC_{0)}$. When RBSEL switches to low to disable the corresponding concurrent refresh mode, CMOS pass gates 622 and 624 turn off and on, respectively, allowing bit $WRAC_0$ to flip and follow node N1, which has been already updated. In conclusion, the bit $WRAC_0$ is flipped at each RBSEL cycle, and acts as the least significant address bit of the row address counter 520. For the remaining address bits ($WRAC_{1\text{-}6}$), CMOS pass gate 622 in element 610_n is coupled to the counter output N4 in the element 610_n-1, where n is an integer from 1 to 6. This allows bits $WRAC_{1, ..., }$ $WRAC_6$ to flip over, respectively, every 2, 4, 8, 16, 32, and 64 RBSEL cycles, creating a seven bit counter to generate a refresh address within each array.

Fig. 7 shows a more detailed bank architecture consisting of array 510, row address counter (RAC 520), and seven bit switching elements 530. The array includes a plurality of memory cells 715 arranged in a two-dimensional matrix as a cell array. The cell array is supported by row decoders 720 and bitline sense amplifiers 725. As previously discussed, the word address for a concurrent refresh mode and a memory access mode are supported by the word address $WRAC_{0, ...,}$ $WRAC_6$ and the word address $WADD_{0, ...,}$ $WADD_6$, respectively. The selection is realized by the seven switching elements 530, each coupling either WADD or WRAC to the row decoder 720. For a memory access mode, the bank select signal BSEL switches to high, activating the NMOS 732 in each switching element530. This couples the word address WADD to the CMOS latch 736. The WADD bit remains even after BSEL is switched to low by CMOS latch 736. For a concurrent refresh mode, the refresh bank select signal RBSEL switches to high, activating NMOS 734 in each switching element 730. This couples the word address WRAC to the CMOS latch 736. Latching the word address in switching element 530 allows the word address $WADD_{0, ...,}$ $WADD_6$ to activate other banks during a multi-

bank operation without waiting for the completion of the corresponding bank operation. Furthermore, during a multi-bank operation, a concurrent refresh operation may be issued for each bank activation cycle. By way of example, the first and second arrays are activated for refresh and memory access operations. Then, without waiting for the completion of the first and second memory array operation, the memory access operation for the third bank is enabled, while concurrently issuing the refresh operation for the forth array.

The wordline WL is activated by row decoder 720. The decoding bits are determined by the address bits transferred from the seven bit switching elements 530, as previously discussed. When the wordline switches to high, the data bits in the cells 715 are read and written through bitline BL. Differential BL pairs are advantageously coupled to bitline sense amplifiers 725, amplifying the small signal read from the cells 715. The sense amplifiers are used for writing back the data bits to the cells (715), which is well know and which, therefore, will not be discussed.

While the invention has been discussed in terms of several preferred embodiments, various alternative and modifications can be devised by those skilled in the art without departing from the invention. Accordingly, the present invention is intended to embrace all such alternatives which fall within the scope of the appended claims.

What is claimed is: